DOCUMENT RESUME

ED 244 561 HE 017 264

AUTHOR Fitt, Paul D.
TITLE Reliable Record Matching for a College Admissions System.
PUB DATE Dec 83
NOTE 13p.; In: Information Resources and the Individual. Proceedings of the CAUSE National Conference (San Francisco, CA, December 11-14, 1983). Boulder, CO, CAUSE, 1984, p. 261-270. For the complete proceedings, see HE 017 245.
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Administration; *College Admission; *Computer Oriented Programs; *Databases; *Data Processing; Higher Education; *Management Information Systems; *Student Records

ABSTRACT

Prospective student data, supplied by various national college testing and student search services, can be matched with existing student records in a college admissions database. Instead of relying on one unique record identifier, such as the student's social security number, a technique has been developed that is based on a number of common data elements that provide a link between the external data source and the existing student record. The matching technique has proven to be highly reliable, has minimized record duplication, has reduced clerical efforts in record maintenance, and can be used with data from various sources. Colleges often have separate prospective student systems related to recruitment that are independent of the applicant processing system that may result in unnecessary record duplication. The record matching technique attempts to reduce this duplication; it incorporates four data elements: social security number, telephone number, high school code, and zip code. The external data is run through a computer program that formats the data into a record form that is compatible to the university's admissions system. Possible duplicates are flagged and processed through a separate program to determine whether they are duplicates. One advantage of the technique is time and money savings in processing achievement and aptitude test results data. A key record format is included. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED244561

RELIABLE RECORD MATCHING FOR A COLLEGE ADMISSIONS SYSTEM

Paul D. Fitt

Administrative Computing Center
St. John's University
Collegeville, Minnesota

This paper describes a successful technique that allows prospective student data, supplied by the various national college testing and student search services, to be matched with existing student records in a college admissions database. Instead of relying on one unique record identifier, such as a social security number which may not always be available for every student, a technique has been developed that is based on a number of commonly occurring data elements, which provides a link between the external data source, and the existing student record. The matching technique has proven to be highly reliable, has minimized record duplication, has reduced clerical efforts in record maintenance, and can be used with data from a variety of sources.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CAUSE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

HE 017 204

Paper presented at the
CAUSE National Conference
(San Francisco, CA, Dec. 11-14, 1983)

Applicant based admissions systems can be characterized by the following qualities:

A) The admissions process was initiated by the student. The completed application form was the originating document that caused the student's record to be entered into the admissions system

B) The student's social security number was used as the unique record identifier in the system

C) Fairly complete personal data about the student was received at one time. Less subsequent record updating per student record was needed

D) The number of student records in the admissions system was predictable from year to year.

PROSPECT/INQUIRY BASED ADMISSIONS SYSTEMS

The last several years in higher education have provided college admissions officers with some of their greatest challenges. With declining birth rates and shrinking applicant pools, admissions officers can no longer wait for students to initiate the admissions process, but must actively seek out and recruit prospective students for their institutions. The need to track students from initial contact through final admissions decision, the demands of market analysis, and the desire to do follow-up contacts are not possible with admissions systems that are based on the receipt of a completed application form.

Many institutions of higher education have developed separate prospective student systems to meet the special needs of student recruitment. These systems exist independently of the applicant processing system. Such dual systems are common among larger colleges and universities where the responsibility for student recruitment is delegated to an office other than the admissions office. Although such dual systems may provide each office with the necessary student information each requires, unnecessary record duplication may result unless the systems are highly coordinated.

A totally integrated system, one that allows both prospective student data and applicant data to co-exist on the same file, is an alternative means to managing admissions data. One major drawback to such an integrated system is that the student's social security number cannot serve as the system's unique record identifier, since it is usually not available at the time of initial student contact. It is this record identifier that allows the student's record to be updated with ACT/SAT test results received from the national testing services.

Increased use of prospective student data acquired from participation in the national student search services, SSS and EOS, presents a similar problem for an integrated system. Although participating institutions can specify selection criteria based on test scores, high school grade point averages, etc., the actual data released to the institution is minimal: name, address, high school code and intended field of study. Incorporation of this data into an integrated system, without duplicating existing records, has proved to be a formidable task.

Prospect/Inquiry based systems can be characterized by the following qualities:

A) Any type of student contact can serve to initialize the student's record into the admissions system, including the use of student search data

DEFINITIONS

Throughout this paper a number of terms will be used that may not be familiar to those persons who are not directly involved in the area of college admissions. For the convenience of the reader these terms are defined below:

ACT	-	A national college admissions test sponsored by the American College Testing Program, Iowa City, Iowa.
Applicant	-	A student who submits an application for admission to an institution of higher education. If accepted the student may enroll at the institution, or cancel his/her application.
Educational Opportunity Service (EOS)	-	One of the national prospective student search services sponsored by the American College Testing program.
Inquiry	-	A student who has expressed an interest in a particular college by taking some action of his/her own, e.g., requesting an application form. An inquiry may apply for admission, be accepted and enroll at the institution.
Prospect	-	A potential inquiry. The student's name has usually been obtained through participation in one of the national student search services. The institution of higher education initiates the contact through an informational mailing.
SAT	-	A national college admissions test sponsored by the College Board, New York, New York.
Student Search Service (SSS)	-	One of the national prospective student search services sponsored by the College Board.

APPLICANT BASED ADMISSIONS SYSTEMS

The development of automated college admissions systems parallels some of the events that have taken place over the last twenty years in higher education. In the mid-1960's, when most larger schools were beginning to automate many of their administrative systems, the words "enrollment decline" were unheard of. In the process of automating the admissions systems of the time, emphasis was placed on duplicating the existing manual systems -- systems that were based on the receipt of a completed application form from a student. A prospective student requested an application form, completed it, and returned it to the school's admissions office for processing. If accepted for admission the student could possibly enroll at the institution, or decided to cancel his/her application. Whatever student recruiting that was done was limited to this pool of applicants.

B) The student's social security number is generally not available at the time of record initiation, and cannot be utilized as a unique record identifier

C) Varying amounts of data can be present at the time of record initiation -- minimally the student's name and address are needed

D) Records are subject to constant updating as new information concerning the student is received

E) The system may contain a high volume of records, perhaps ten times the usual number of applicant records.

Since its beginning on second generation computing equipment in the late 1960's, and after a conversion to more modern hardware in 1978, the admissions system at Saint John's University has been an inquiry based system. A student record was initiated on the system upon initial student contact, and it was possible to track students from their initial contact, through their final admissions decision.

During the last several years the university has participated in the Student Search Service of the College Board, and recently the new Educational Opportunity Service of the American College Testing Program. Typically, 20,000+ pre-selected prospective student names and addresses are received from Student Search on gummed labels and are used for an informational mailing. When the student responded to the mailing by returning a response card, i.e., the student inquiried for additional information, a record for him was initiated on the admissions system, and an application form sent out. Although it was quite easy to follow-up on those students who inquired, but did not apply, it was not possible to conveniently follow-up on those students who were sent the original mailing, but who did not return their response cards.

For the last two years student search data has been received in the form of a computer tape file that allows the university to produce its own mailing labels, and also provides the opportunity to "reuse" the data for follow-up mailings. Having the student search data available in machine readable form proved to be a mixed blessing for both the admissions office and computing center staffs. Follow-up mailings to non-responding students were a possibility, but the inconvenience of separating respondents from non-respondents still existed. It was also not possible to directly add the data to the admissions system for fear of creating a number of duplicate records for the same student. This would only add to the confusion in subsequent mailings with many students receiving two or three promotional pieces. The computing center staff was also beginning to be inundated with requests for reports from the student search tape. Although such reports were possible to produce, it became increasingly apparent that, unless the data from various external sources could be consolidated into the current admissions system, separate "systems" would begin to develop around each of about six external data sources.

GOALS OF THE PROJECT

The challenge at hand then was to develop a record matching technique that would allow data from various external sources to be reliably matched with that currently on the admissions system. To be considered successful the matching technique would have to meet the following goals:

A) Be based on data elements common to both existing student records and those from external sources

B) Work with data from various sources; both highly data intensive sources such as the ACT/SAT test reports, and such minimal data sources as Student Search and Educational Opportunity records

C) Significantly reduce manual duplicate record checking and record updating

D) Minimize the generation of duplicate student records

E) Be highly reliable.

DATA UNIQUENESS

Over a year's time a number of experiments were performed with a test file of 25,000+ records to determine what combinations of data elements, e.g., first and last name, telephone number, zipcode, etc., could produce an unique and reproducible record identifier that could serve to match external source records with those already in the admissions file. The experimental results indicated that no totally unique identifier could be produced that would serve a useful purpose. Unique identifiers could be produced from the test data, but required so many data elements that they were not reproducible from the data contained in the admission system records.

The experimentation did, however, produce some indication as to the level of uniqueness of various data elements. Social security numbers (present in some records only) and telephone numbers proved to have a high degree of uniqueness. Some student names, e.g., Zymond Zyvifkovitch, tend to be unique, where as others, e.g., John Smith are not.

KEY INTERFACE TECHNIQUE

The record matching technique that was finally settled upon was one that incorporated the various levels of uniqueness of four data elements into a hierarchy of uniqueness. The four data elements that were selected based on the experimental results were: social security number, telephone number, high school code, and zipcode.

Each of the data elements is used together with parts of the student's first and last name to produce a key record that is loaded to an intermediate key file (figures 1, 2). It is this key file that serves as an interface between student records in the external data source, and those in the admissions student master file (figure 3).

Portions of the student's name were incorporated into the key record in order to give those data elements with lower levels of uniqueness, e.g., high school code, zipcode, a higher degree of uniqueness. This procedure guaranteed that data from some sources, i.e., SSS, EOS, which do not contain social security and telephone numbers, would be processed just as reliably as those records that do, i.e., ACT/SAT test report tapes.

Before processing data from any external source the intermediate key file is generated from data contained in the admissions student master file. During this processing step 1-4 key records per student record are produced depending upon the presence of the four data elements in the record. An active admissions master file of 31,000+ records produces on the average 2.2 key records per student with a 99.8% of the key records being unique. The majority of keys produced tend to be high school code and zipcode keys.

EXTERNAL DATA SOURCE PROCESSING

Regardless of the external data source two facts are known: some of the student records will be duplicates of those already on the university's admissions system, and some will be new records. The successful separation of records into these two categories requires a two step process.

In the initial step (figure 4) the external data is run through a computer program that formats the data into a record format that is compatible to the university's admissions system. Due to the different record formats used by each of the external data sources, a separate formating program is used for each external data source. During the formating process a determination is made as to whether an external record is "new" or a "possible" duplicate. This determination is accomplished by repeating the key generation process described above for each of the external data records. The 1-4 key records are generated and an attempt is made to write these new key records to the previously generated key file. If a successful write occurs, the external student record is considered new, since no other student record on the admissions system has produced the same key record. New student records are written to a file to await addition to the admissions student master file by a separate computer program. The determination of the "new" student record status has proven to have a high degree of reliability, 98-100%, depending on the external data source.

The "possible duplicate" status is determined when a key write attempt has failed -- a student record on the admissions system has already produced a key similar to that of the external data record. The records are considered <u>possible</u> duplicates at this time since some identical lower level keys can be produced for two entirely different students, i.e., students with common names living in the same zipcode area. Possible duplicate records are written to a file to await further processing by a separate update program.

DUPLICATE RECORD PROCESSING

The second phase of record processing involves the <u>possible</u> duplicate record file produced by the formating program (figure 5). The possible duplicate file is processed by a computer program that determines whether or not the record can be considered an <u>actual</u> duplicate record. This determination is made by comparing the possible duplicate record with its match record on the admission system. If social security numbers are present in both records these are compared, and if equal, a positive match is assumed, and the existing student record is updated with any new information contained in the external record. If social security numbers are not available for either of the two records, then the student names are examined, and if the names agree in both records, a match is assumed and the existing record updated.

In those cases where positive matches do not occur, both records are outputted to an exception report. At this point there is a possibility that the two records belong to the same student, but, because of name variations (Tom/Thomas, Bill/William, etc.) a positive match was not obtained. The exception report is reviewed by a clerk who makes the final determination as to whether or not the records are duplicates. The clerk either updates a duplicate record, or adds a new student record to the system using an on-line add/update program. During the last several months that the record matching technique has been used 2,974 duplicate records were processed with only 173 (5.8%) needing manual reconciliation by a clerk. Most of the records needing reconciliation tended to be actual duplicates due to name variations.

Another function of the updating program is the verification of a number of existing data elements common to both the external data record and the admissions master file record. For example, if both contained telephone numbers no updating would have taken place, but a check would be made to ascertain if the numbers were identical. Any discrepancies are outputted to the exception report for reconciliation by a clerk.

ADVANTAGES/DISADVANTAGES

Since the record matching technique has been implemented a number of distinct advantages have become evident:

A) Savings in both time and money in the processing of ACT/SAT test results data

B) Expanded student recruitment pool integrated into a unified system

C) Expanded use of student search data with more follow-up opportunities

D) Ten-fold increase in admission records with no increase in clerical staff

E) Less manual record processing, and less data entry errors

F) Routine processing of all external data sources

G) Market analysis statistics are available for all prospective students, not just for those who responded to an initial mailing.

Although the results of the technique has been overwhelmingly positive, a few disadvantages have resulted due mainly to the ten-fold increase in the number of prospective student records on the admission file. These include:

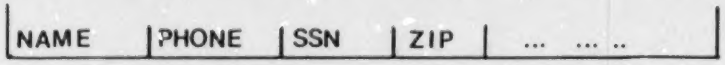
A) Increased computer storage space/cost

B) Longer processing time for routine selection programs

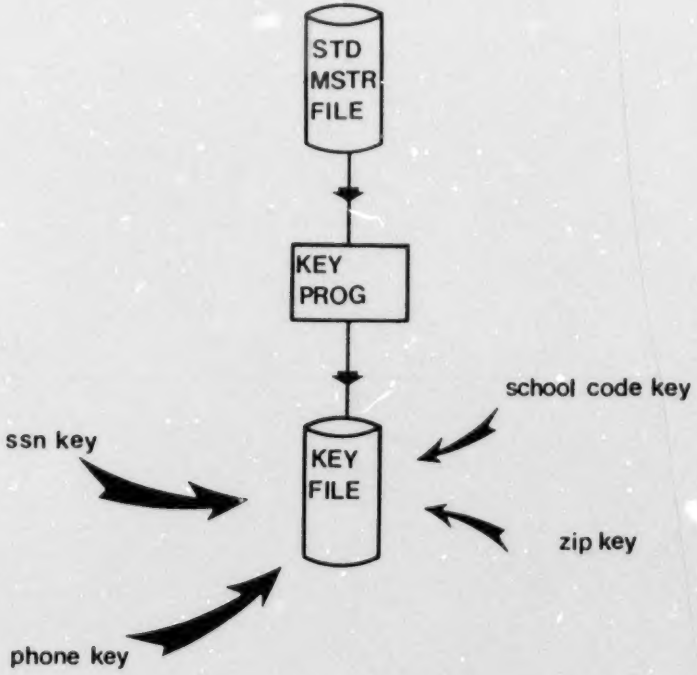
C) Need to use selective reports rather than complete file reports.

COST BENEFITS

During the last several months of operation over 31,000 records were processed using the described matching technique. If these records were processed using manual data entry techniques over 2,600 hours of labor would have been required. Assuming work study wages of $3.75 per hour over $9,800 would have been spent in accomplishing this task.



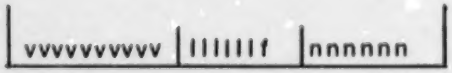
STUDENT RECORD



1–4 KEYS PER STUDENT

Figure 1

KEY RECORD FORMAT



vv···vv	SSN or PHONE or SCHOOL CODE or ZIP
llllllf	6 LAST NAME 1 FIRST
nnnnnn	ID NUMBER

Figure 2

INITIAL FILE PROCESSING

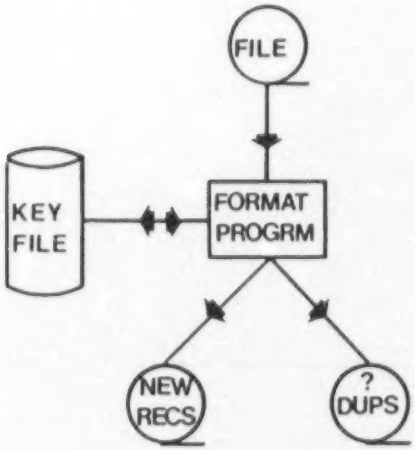


Figure 4

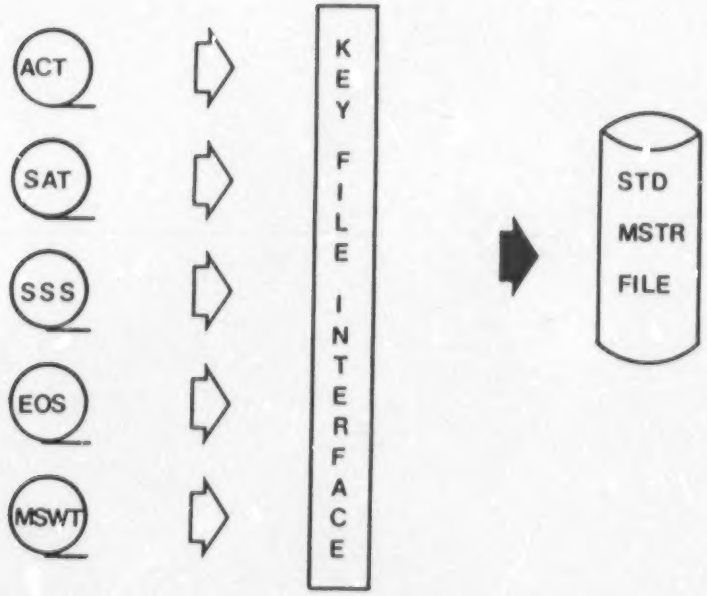
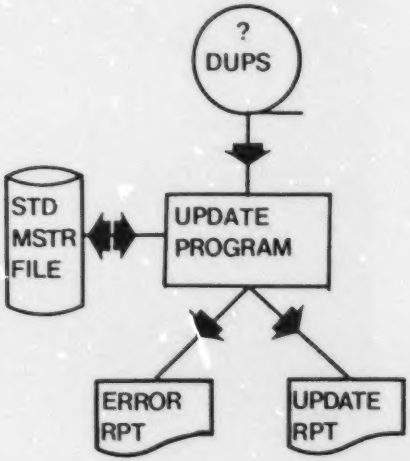


Figure 3

DUPLICATE RECORD PROCESSING



VERIFIES:

- school code
- year of entry
- term of entry
- phone no.
- etc.

Figure 5